Published online 4 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D939–D944***
doi:10.1093/nar/gkq957

VnD: a structure-centric database of disease-related SNPs and drugs

Jin Ok Yang[1], Sangho Oh[1], Gunhwan Ko[1], Seong-Jin Park[1], Woo-Yeon Kim[1], Byungwook Lee[1,*] and Sanghyuk Lee[1,2,*]

[1]Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Daejeon 305–806 and [2]Ewha Research Center for Systems Biology, Division of Life and Pharmaceutical Sciences, Ewha Womans University, Seoul 120–750, Korea

Received August 15, 2010; Accepted September 30, 2010

ABSTRACT

Numerous genetic variations have been found to be related to human diseases. Significant portion of those affect the drug response as well by changing the protein structure and function. Therefore, it is crucial to understand the trilateral relationship among genomic variations, diseases and drugs. We present the variations and drugs (VnD), a consolidated database containing information on diseases, related genes and genetic variations, protein structures and drug information. VnD was built in three steps. First, we integrated various resources systematically to deduce catalogs of disease-related genes, single nucleotide polymorphisms (SNPs), protein mutations and relevant drugs. VnD contains 137 195 disease-related gene records (13 940 distinct genes) and 16 586 genetic variation records (1790 distinct variations). Next, we carried out structure modeling and docking simulation for wild-type and mutant proteins to examine the structural and functional consequences of non-synonymous SNPs in the drug-related genes. Conformational changes in 590 wild-type and 4437 mutant proteins from drug-related genes were included in our database. Finally, we investigated the structural and biochemical properties relevant to drug binding such as the distribution of SNPs in proximal protein pockets, thermo-chemical stability, interactions with drugs and physico-chemical properties. The VnD database, available at http://vnd.kobic.re.kr:8080/VnD/ or vandd.org, would be a useful platform for researchers studying the underlying mechanism for association among genetic variations, diseases and drugs.

INTRODUCTION

Discovering genetic factors affecting disorders or diseases is crucial for understanding the pathogenesis, diagnosis and treatment of human diseases. Previous studies indicate that single nucleotide polymorphisms (SNPs) are the most common type of DNA sequence variation found in human genome, accounting for at least 1% of the genetic differences between individuals (1,2). In particular, non-synonymous SNPs (nsSNPs) in the coding region of a gene can alter the function or structure of protein by changing amino acids or introducing a premature stop codon (3). Conformational changes in these proteins are major targets for drug development. Indeed, drug response to these genetic variations has emerged to be a major subject in the field of pharmacogenomics with the combined use of genetics and functional genomics data. Information on SNPs and structural changes in disease-related proteins is thus important in biomedical studies, diagnostics and drug development (4).

Both public and commercial databases exist to provide information on relationship between genetic variants and drug targets. Such public efforts are represented by GenoWatch (5), IDBD (6), DrugBank (7) and SuperDrug (8). The GenoWatch and IDBD databases contain information about specific diseases and a browser for disease–gene association studies. DrugBank contains details on drugs such as drug target and action, and SuperDrug provides three-dimensional (3D) structures and conformers of drugs. Although each database has its own objectives, they provide information of limited scope such as disease-associated genes, genetic variations

*To whom correspondence should be addressed. Tel: +82 42 879 8511; Fax: +82 42 879 8519; Email: bulee@kribb.re.kr
Correspondence may also be addressed to Sanghyuk Lee. Tel: +82 42 879 8500; Fax: +82 42 879 8519; Email: sanghyuk@kribb.re.kr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

and drugs or 3D structural models of drugs. The commercial sector, led by the World Drug Index (9), Chemistry, Manufacturing and Controls (CMC) (10) and the MDL Drug Data Report (11), provides more comprehensive coverage. However, they are usually very expensive and accessible only by private commercial entities.

Protein structure modeling and docking simulations require computational power and experts. To our knowledge, no public resource is available to cover the structural aspect of disease proteins taking their genetic variations into account. Furthermore, effect of genetic variations on docking with drugs would be valuable information for drug development.

Here, we present a database, variations and drugs (VnD), which provides comprehensive information on diseases-related genes, their genetic variations, protein structure modeling and docking simulations. More specifically, available information is as follows: (i) a comprehensive catalog of disease-related genes, proteins and drugs; (ii) structural changes caused by nsSNPs in disease-related genes; (iii) their consequences in drug binding using docking simulation such as AutoDock (12), Dock (13) and Fred (14) programs; (iv) distribution of nsSNPs near the structural pockets in disease-related proteins; and (v) functional effects of SNPs known to be related to common diseases from association studies.

DATABASE DESIGN AND CONTENT

To build the VnD database, we developed an automatic pipeline as shown in Figure 1. It consists of three main steps: (i) collection of disease-related genetic variations and proteins from public disease databases using ontology-based unification of disease terms, (ii) structure modeling for both wild-type and nsSNP mutant proteins and (iii) analysis of protein structures and identification of potential drug binding sites.

Collection of genetic variations associated with diseases

Disease term unification. We extracted disease terms from two disease databases: OMIM (15) and GAD (16). Unfortunately, these databases use highly inconsistent terminology to describe the same disease. For example, 141 slightly different disease descriptions exist for 'Parkinson's disease'. Therefore, we used the Unified Medical Language System (UMLS) (17), which contains medical subject headings (MeSHs) and clinical terms from the systematized nomenclature of medicine to standardize the disease terms. The disease terms in OMIM and GAD were mapped on the concept unique identifier (CUI) in UMLS (18) taking disease synonyms into consideration. Through this unification procedure, we obtained 36 109 disease terms, which were then mapped to 3898 CUIs (see Supplementary Table S1 for statistics).

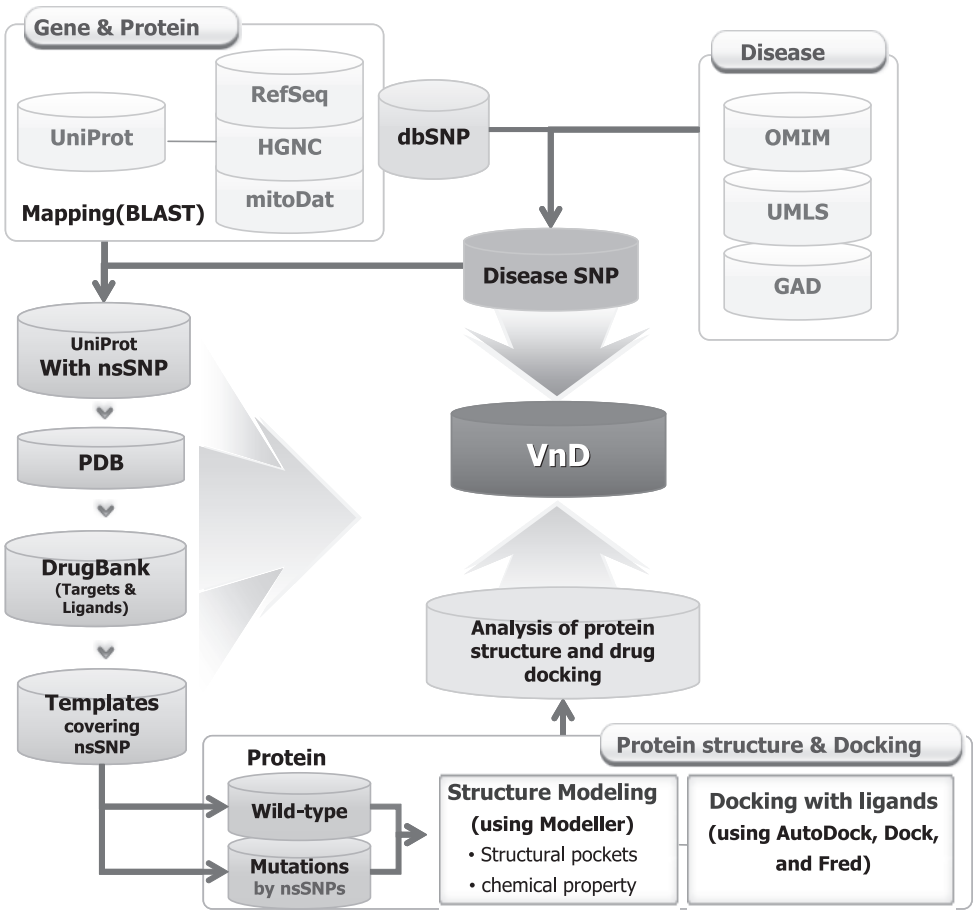

Figure 1. The workflow of the VnD database.

Collection of disease-related genes. We extracted the candidate genes associated with diseases or disorders based on genomic positions and gene names. To cover the name space of disease-related genes, we extracted 40 234 gene names from the HUGO Gene Nomenclature Committee (HGNC) (19) and the NCBI Gene database (20). We integrated the genome annotation data as well from various sources: NCBI's Entrez Gene (20), RefSeq mRNA from the UCSC table track (21) and protein information from UniProt (22). RefSeq mRNAs were mapped to genes, and 85 510 proteins were linked to genes using the BLAST (23) search. Ultimately, we obtained 13 940 disease-related genes and 10 883 disease-related UniProt proteins (Supplementary Table S2).

Collection of disease-related genetic variations. As a source of genetic variations, we used the databases of dbSNP (24) and JSNP (25). Representative SNPs were mapped onto genes and proteins based on the SNP loci and identifier (rs numbers). Total number of representative SNPs was over 14.5-million. The number of SNPs in the genic region was 5 766 017, where 91 038 SNPs were non-synonymous. Among the amino acid changes caused by nsSNPs, changes in glycine affect protein structure and function most dramatically. Glycines at certain position are strongly conserved evolutionarily due to the size restriction in protein structure. Mutations at such sites would affect the structure and function of the protein significantly (26). We examined the mutation spectrum of amino acids changes caused by nsSNPs (see the website for detailed result), and found a total of 5034 (6.2%) glycine changes due to nsSNPs. In an effort to predict the functional aspects of these nsSNPs, we have analyzed the disease risk for 91 038 nsSNPs using polymorphism phenotyping (PolyPhen) (27).

Modeling structural changes in protein due to genetic variations

To predict structural changes in the drug-related proteins, we have selected 2486 proteins out of 10 883 disease-related proteins that showed sequence similarity over 95% identity with the drug target sequences in the DrugBank database. Search for structural templates for homology modeling was carried out using the BLASTP and PSI-BLAST methods with the minimum percent identity of 60% for the proteins in the PDB structure database (28). We filtered out templates with less than 100 amino acids. This procedure produced the structural templates for 601 drug-related proteins.

Among the candidate templates that covered the nsSNP positions, we selected the template with the highest identity as the primary template. Then the secondary-structure alignment, which is the input for Modeller, was carried out using the local PSI-Pred. Next, we performed 3D structural modeling for drug-related proteins using Modeller (version 9v7) with a single template. Modeller automatically constructs an all-atom 3D model using one or more alignments between the query sequence and the homologous protein sequences of known structure (29). Finally, we determined the best 3D structural model based on the highest stability energy score (z-score).

To examine the structural changes due to amino acid substitution, we generated 4020 mutant proteins at known nsSNP sites. Structural modeling for mutant proteins was carried out in a similar fashion using the same template as the wild-type proteins (see Supplementary Figure S1 for more details). In summary, we constructed 3D structural models for 590 wild-type proteins and 4437 mutant proteins from 538 proteins considering the disease-related nsSNPs (see Supplementary Table S3).

Analysis of protein structural changes and docking simulation

We have analyzed the difference in structural stability between wild-type and mutant proteins. The $\Delta\Delta G$ score of each mutant versus wild-type proteins was calculated using the I-mutant program (version 2.0). This program calculates the free energy difference to estimate the stability change due to mutations (30). Positive $\Delta\Delta G$ scores indicate an increased stability. Large values for $\Delta\Delta G$ (absolute value >1) may indicate significant structural changes, which could affect the drug binding by changing the pocket size or shape (30,31).

Previous studies have reported that protein functions are highly dependent on physical, chemical and geometric features of pockets on the surface of the protein (32,33). Changes in pocket size or stability due to nsSNPs can affect the interactions between target proteins and ligands. Thus, nsSNPs close to the structural pockets are likely to have deleterious effects to be the cause of disease (34) or differences in drug metabolism. To identify the SNP distribution near the pockets, we analyzed the pockets in protein structure using the LIGSITE, which calculates the pocket size and potential ligand-binding sites by the protein–solvent–protein method (35). We examined the pocket sizes up to 10 000 $Å^3$, allowing overlap of maximum three pockets. Most pockets were found to be in the range between 20 and 4000 $Å^3$. More than 50% of nsSNPs were located inside the first two largest pockets.

We also calculated the distances between nsSNPs sites and the structural pockets. It was found that 767 (17%), 2176 (49%) and 3192 (71%) nsSNPs were located within pockets, 5 Å from pockets and 10 Å from pockets, respectively. Because atoms within ~5–6 Å are able to interact with each other (36), these SNPs can influence interactions between the target protein and ligands.

In an effort to provide the structural picture of drug binding, we performed the docking simulations between the drug with the target and the mutant proteins. Three public programs—AutoDock (version 4.0), Dock (version 6.0) and Fred (version 2.0), were used with the default options and we obtained 981 docking results.

WEB INTERFACE

The VnD web page supports four types of search for user convenience—protein, gene, SNP identifier and disease. Example outputs from the VnD are shown in Figure 2.

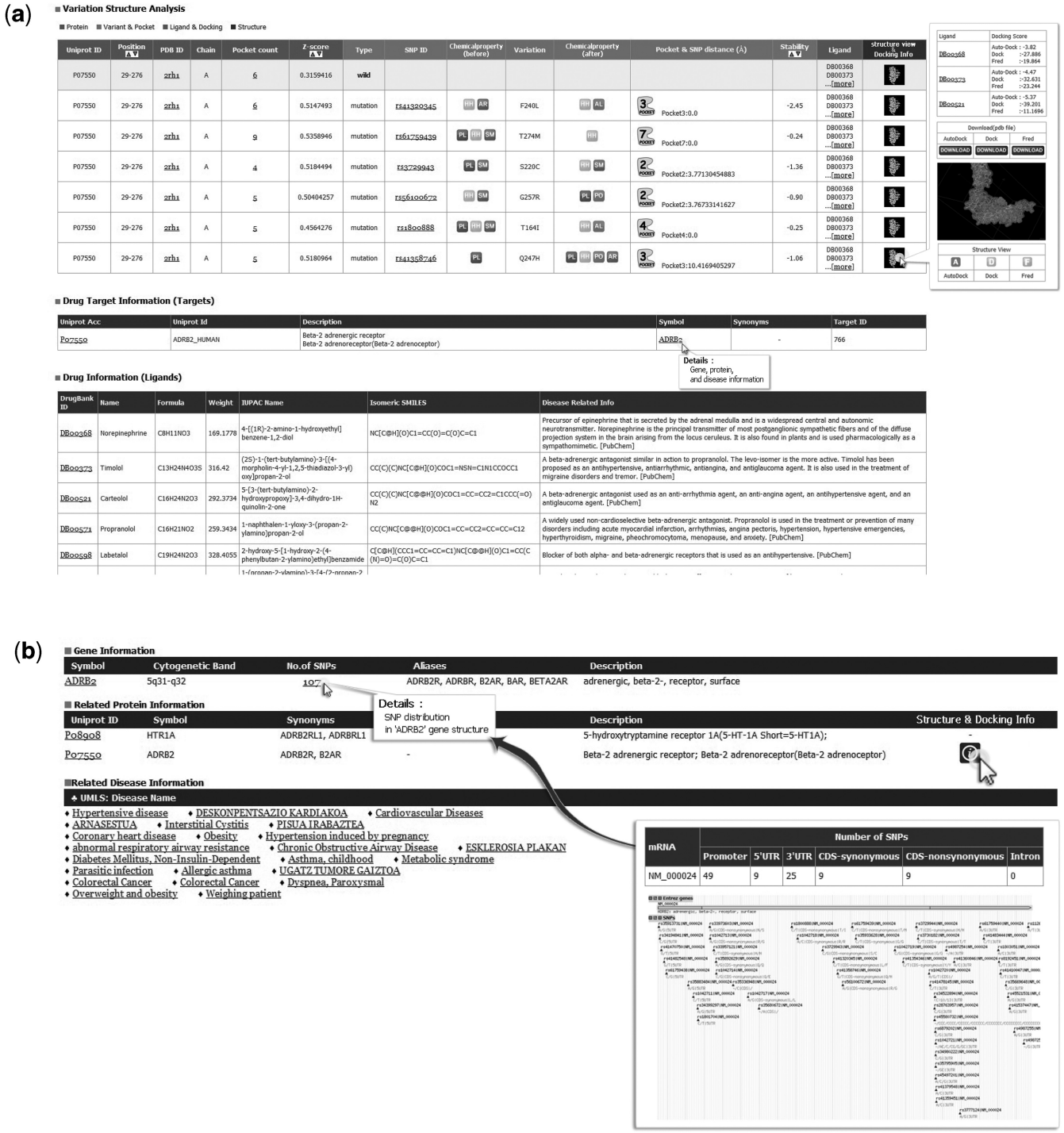

Figure 2. Query table results and graphic viewer. (**a**) The server displays information on structures of wild-type and mutant proteins and drug docking found as results for a protein query. The 'distance between pocket and SNP' column indicates a pocket located close to a SNP and the distance between a pocket and SNP. The 'structure view' column provides the structures of the wild-type and SNP mutant proteins and docking with ligands. (**b**) Clicking on the gene symbol in the results table (a) allows the user to see the SNP distribution located in the gene structure, protein and disease information.

In the protein menu, users can input a protein ID (UniProt or PDB) and obtain its structural properties, changes by nsSNP(s) and ligand docking information from three public programs. When the number of pockets is clicked, users can observe information about the pockets located in the target protein. Clicking the 'structure view' link allows users to observe the protein structure with the Jmol visualization software (http://jmol.sourceforge.net) and download its 3D structural information.

In the Gene menu, users are able to view the SNP distribution and location in the query gene, related protein information and the relevant disease information as shown in Figure 2b. By clicking the 'No. of SNPs' in the 'Gene Information' table, information on transcripts and SNP markers is displayed in the GMOD genome browser (37). This would facilitate the recognition of disease-related genetic features such as SNPs within the promoter region or near the splice sites (38).

In the SNP menu, users can obtain detailed information on the SNP including the disease risk estimated from PolyPhen. One can also explore the structural changes in related proteins if the query SNP is nonsynonymous. In addition, the VnD web interface provides a tree view of the disease terms in the UMLS concepts. Currently, the tree view of disease terms consists of 23 top disease terms having an average of five or six sub classes.

To demonstrate the usefulness of the VnD server, we provide the β-2 adrenergic receptor protein (P07550) as an example case. The output pages in Figure 2 can be classified in three categories: (i) physical properties and conformational changes due to nsSNPs in the query protein; (ii) query protein and drug target protein information and (iii) drug ligands and side effect information. Specifically, this query protein is associated with obesity, diabetes, parasitic infection and asthma. The 3D structure and the number of functional sites in the protein are also available in the output. Furthermore, changes in chemical and physical properties such as energy stability caused by six disease-related nsSNPs are also shown. Remarkably, one of the nsSNPs (rs56100672) causes an amino acid substitution (G257R) that changes a small, hydrophobic residue glycine into a polar, bulky, and positively charged residue. The 3D structural models for wild-type and mutant proteins are shown in Supplementary Figure S2. It shows that the pocket size is reduced significantly from 214 to 170 $Å^3$. This size reduction and changes in the pocket shape may have some relationship with the disease and drug susceptibility which need further studies. Therefore, users can observe how the disruption of the surface pocket may affect the protein function and explore its relationship with the molecular causes of a disease or different drug susceptibilities among individuals.

The VnD database server is composed of a web interface and a MySQL (version 5.0.45) database management system. The web interface is implemented in static HTML pages, JSP and Java (version 1.6.0_20). MySQL is used to store the disease-related and drug information.

CONCLUSION

We have constructed a comprehensive database that provides information on genetic variations of disease-related genes and their structural and functional consequences in the aspect of drug target proteins. The effects of non-synonymous SNPs in disease- and drug-related genes were of special focus. We carried out diverse analyses for wild-type and mutant proteins, which include homology modeling, docking, disease risk assessment and analysis on pockets and structural features. Results from all these analyses were integrated into a user-friendly website that would facilitate a mechanistic understanding of trilateral relationships among the genetic variations, diseases and drugs.

The number of disease- and drug-related genes is rapidly increasing partly due to the recent advances in the genome-wide association studies (GWAS). The list of disease-related mutations is expanding as well, as the next-generation sequencing (NGS) techniques become a routine practice. The VnD database will continue to serve as the platform site to explore the relationship between genetic variations and drug effects based on structural modeling and docking simulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors thank Ms.Eujin Kwak for editing the web figures.

FUNDING

Korea Research Institute of Bioscience and Biotechnology (KRIBB) Research Initiative Program and 'Systems Biology Infrastructure Establishment Grant' provided by Gwangju Institute of Science & Technology in 2010 through Ewha Research Center for Systems Biology (ERCSB). Funding for open access charge: KRIBB Research Initiative Program.

Conflict of interest statement. None declared.

REFERENCES

1. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
2. Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al*. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet*., **22**, 231–238.
3. Yang,J.O., Hwang,S., Oh,J., Bhak,J. and Sohn,T.K. (2008) An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases. *BMC Bioinformatics*, **9(Suppl. 12)**, S19.
4. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
5. Chen,Y.H., Liu,C.K., Chang,S.C., Lin,Y.J., Tsai,M.F., Chen,Y.T. and Yao,A. (2008) GenoWatch: a disease gene mining browser for association study. *Nucleic Acids Res*., **36**, W336–W340.
6. Yang,I.S., Ryu,C., Cho,K.J., Kim,J.K., Ong,S.H., Mitchell,W.P., Kim,B.S., Oh,H.B. and Kim,K.H. (2008) IDBD: infectious disease biomarker database. *Nucleic Acids Res*., **36**, D455–D460.
7. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*., **36**, D901–D906.
8. Goede,A., Dunkel,M., Mester,N., Frommel,C. and Preissner,R. (2005) SuperDrug: a conformational drug database. *Bioinformatics*, **21**, 1751–1753.

9. Voigt,J.H., Bienfait,B., Wang,S. and Nicklaus,M.C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, **41**, 702–712.
10. Sachs,H. (1997) Quality control by the Society of hair testing. *Forensic Sci. Int.*, **84**, 145–150.
11. Sheridan,R.P. and Shpungin,J. (2004) Calculating similarities between biological activities in the MDL Drug Data Report database. *J. Chem. Inf. Comput. Sci.*, **44**, 727–740.
12. Huey,R., Morris,G.M., Olson,A.J. and Goodsell,D.S. (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, **28**, 1145–1152.
13. Bikadi,Z. and Hazai,E. (2009) Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *J. Cheminformatics*, **1**, 15.
14. McGaughey,G.B., Sheridan,R.P., Bayly,C.I., Culberson,J.C., Kreatsoulas,C., Lindsley,S., Maiorov,V., Truchon,J.F. and Cornell,W.D. (2007) Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inform. Model.*, **47**, 1504–1519.
15. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
16. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nature Genet.*, **36**, 431–432.
17. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.
18. Bae,J.S., Cheong,H.S., Kim,J.O., Lee,S.O., Kim,E.M., Lee,H.W., Kim,S., Kim,J.W., Cui,T., Inoue,I. *et al.* (2008) Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. *Biochem. Biophys. Res. Commun.*, **373**, 593–596.
19. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
20. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
21. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E., Siepel,A. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
22. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol. Biol.*, **406**, 89–112.
23. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
24. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
25. Hirakawa,M., Tanaka,T., Hashimoto,Y., Kuroda,M., Takagi,T. and Nakamura,Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.*, **30**, 158–162.
26. Parrini,C., Taddei,N., Ramazzotti,M., Degl'Innocenti,D., Ramponi,G., Dobson,C.M. and Chiti,F. (2005) Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure*, **13**, 1143–1151.
27. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nature Methods*, **7**, 248–249.
28. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
29. John,B. and Sali,A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.*, **31**, 3982–3992.
30. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
31. Carey,P.R. and Dong,J. (2004) Following ligand binding and ligand reactions in proteins via Raman crystallography. *Biochemistry*, **43**, 8885–8893.
32. Liu,Z.P., Wu,L.Y., Wang,Y., Chen,L. and Zhang,X.S. (2007) Predicting gene ontology functions from protein's regional surface structures. *BMC Bioinformatics*, **8**, 475.
33. Laskowski,R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.
34. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, **32**, D520–D522.
35. Hendlich,M., Rippmann,F. and Barnickel,G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model*, **15**, 359–363, 389.
36. Stitziel,N.O., Tseng,Y.Y., Pervouchine,D., Goddeau,D., Kasif,S. and Liang,J. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol.*, **327**, 1021–1030.
37. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
38. Abu,A., Frydman,M., Marek,D., Pras,E., Stolovitch,C., Aviram-Goldring,A., Rienstein,S., Reznik-Wolf,H. and Pras,E. (2006) Mapping of a gene causing brittle cornea syndrome in Tunisian jews to 16q24. *Invest. Ophthalmol. Vis. Sci.*, **47**, 5283–5287.